WZB

Text-mining international politics

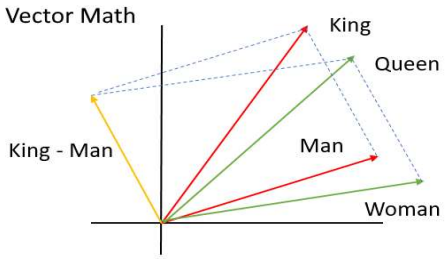
Session 6
Advanced methods and outlook

1

WZB

Word vector models / Word embeddings

- Can you solve?
  'king' – 'man' + 'woman'



- Words referring to similar concepts should have similar vectors
  –> *semantic meaning*

Figure source: https://blogs.mathworks.com/loren/2017/09/21/math-with-words-word-embeddings-with-matlab-and-text-analytics-toolbox/

2

WZB

Word vector models / Word embeddings

- 'Embed' words into a vector space model based on co-occurrence with other words in document-frequency matrix
  - Usually ~ 300-1.000 dimensions; context windows of ~ 100 terms
  - Trained on large corpora ; prominent algorithms are *GloVe* (Pennington et al 2014) and *word2vec* (Mikolov et al. 2013)
  - Pre-trained models but also free implementations (e.g. in R) exist
  - Note: Results do not generalize beyond the training corpus!
- Possible applications for international politics and its analysis
  - Expansion of dictionary-based analyses based on 'semantic synonyms'
  - Distance of terms across different speakers (e.g. 'China' and 'trade')
  - Varying meanings and/connotations of terms across time (or time-sliced corpora)
  - ...
- *See Spirling/Rodriguez via Github*

WZB

Text distance / similarity

- Similarity/distance along the 'bag of words'
  - Jaccard: shared terms/unique terms
  - Euclidian: square root of summed squared diffs across terms
  - Cosine: 'angle' between the two word vectors
- Similarity/distance respecting character or term order
  - N-gram similarity: # of shared n-grams / # of unique n-grams
  - Longest common substring
  - Minimum edit distance algorithms: how many edit operations does it take at least to transform one text into the other?
    - Damerau-Levenshtein: insertion, deletion, substitution, or adjacent transposition of individual characters or terms
    - DocuToads (Hermannsson/Cross 2017 *EUP*): similar, but 'punishes' less for transposition of large text chunks

WZB

Text distance / similarity

- Developed for spelling checkers and plagiarism detection, but very promising for the study of international politics …
  - Diffusion of policy ideas
  - Similarity of international treaties and agreements
  - Change across consecutive negotiation documents
  - …

- But choose your similarity/distance measure wisely!

  "We should focus on environmental protection and take care about economic growth as well."

  "We should focus on economic growth and take care about environmental protection as well."

5

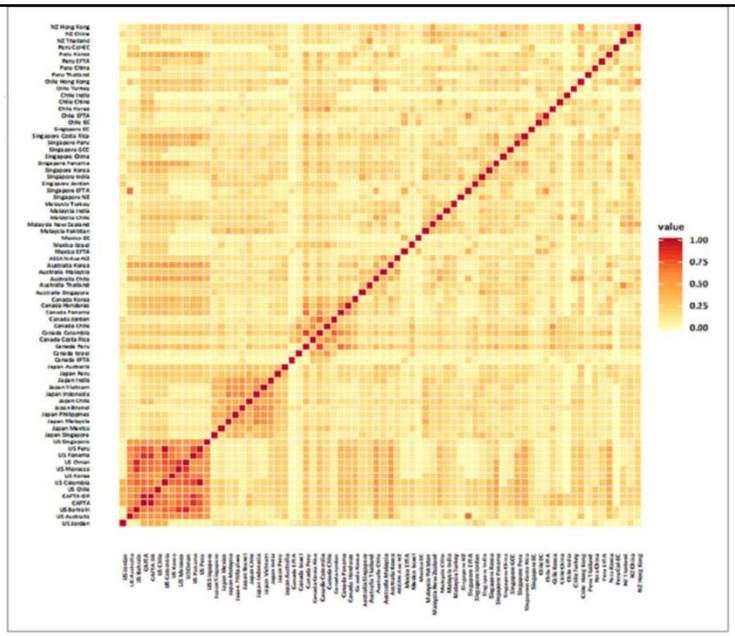


Figure 1. Heat map of text similarity among Trans-Pacific Partnership members' preferential trade agreements, 1995–2015.

Source: Allee/Lugg (2016, *Research & Politics*)

6

WZB

Who wrote the Trans-Pacific Partnership (TPP)?

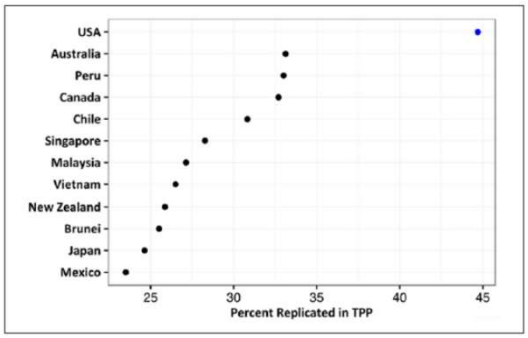


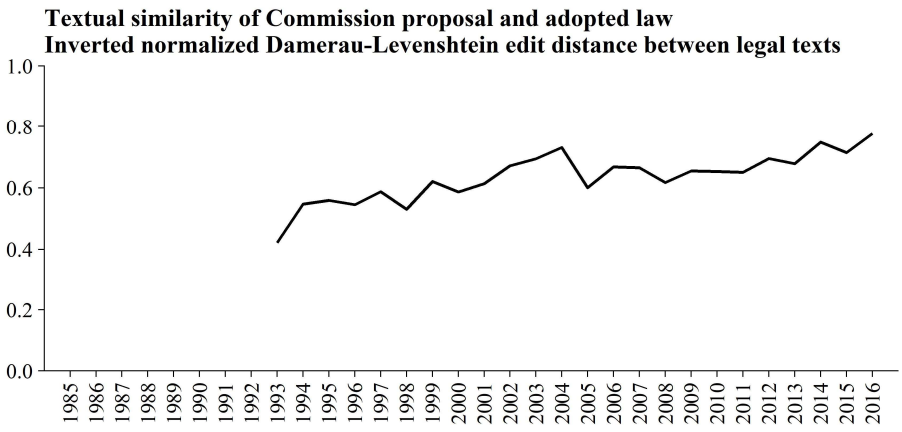
Figure 3. Average amount of past preferential trade agreement text replicated in the Trans-Pacific Partnership, by country.

Source: Allee/Lugg (2016, *Research & Politics*) 7

7

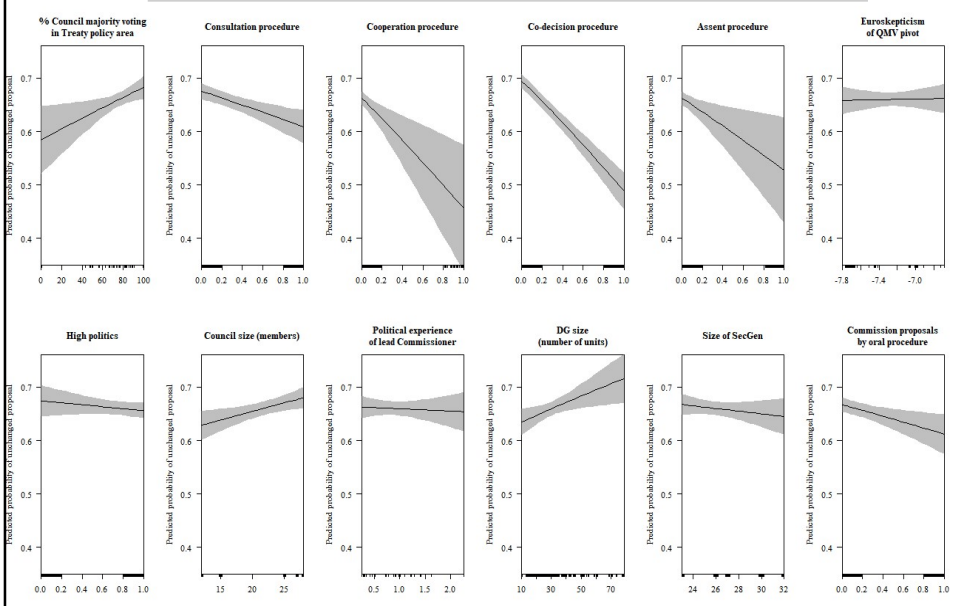
WZB

Legislative agenda-setting success of the European Commission



Source: Rauh (2018, *Researchgate*) 8

8



Source: Rauh (2018, *Researchgate*)

9

WZB

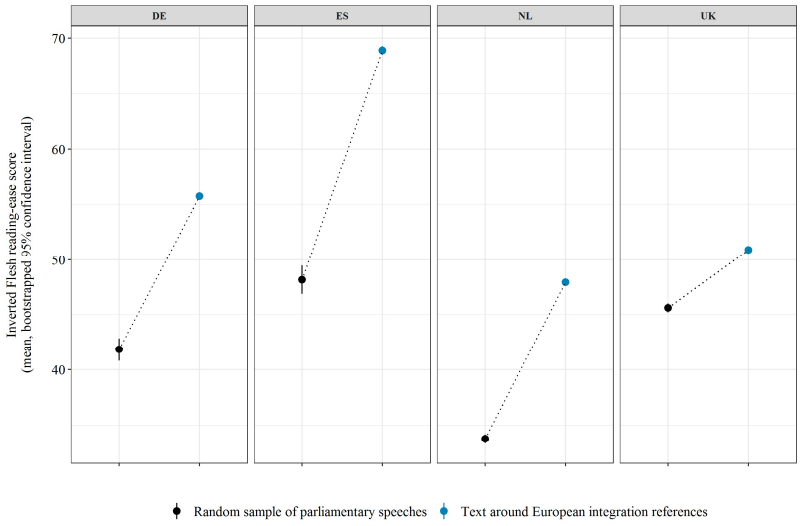
Language complexity

- **Cognitive mobilization required to understand a text (message)**
  - *Flesh–Kincaid reading ease score*: weighted index of term and sentence length (interpretable along U.S. education levels)
  - *LIX readability test* (Björnsson 1968):
    Index of overall number of terms, number of (sub-)clauses, and number of particularly long terms (language specific weights)
  - '*Political sophistication*' (Benoit, Munger, Spirling 2019, PA):
    Includes a.o. contemporaneous word rarity retrieved from their relative frequency in the annual Google Books corpus
- **Very interesting for analysing international politics**
  - Quality of political communication
  - Complexity of policies
    (relevant variable in bargaining and delegation theory, e.g.)
  - *Your ideas here …*

10

WZB

Is Europe that hard to explain?

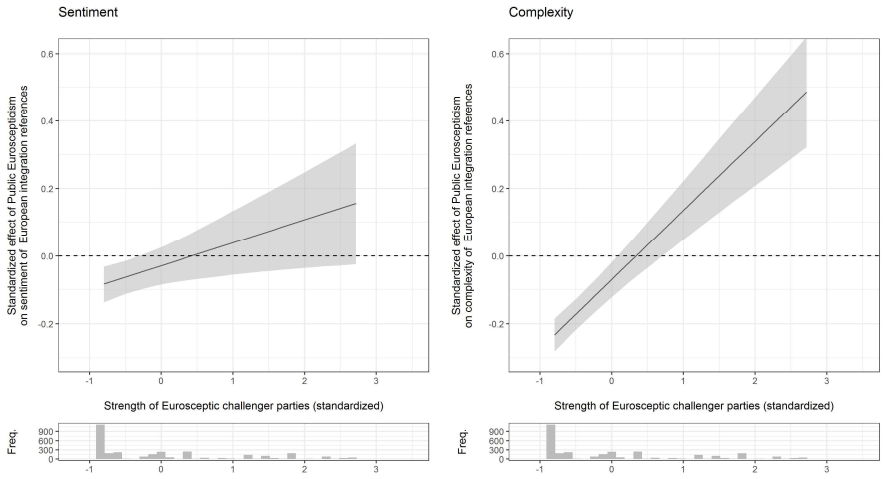


Source: Rauh (2019, mimeo), national plenary debates from 1995-2013 (drawn from ParlSpeech) 11

11

WZB

National leaders' EU communication under varying levels of public/partisan Euroscepticism



Source: Rauh / Bes / Schonnvelde (2019, EJPR, forthcoming) 12

12

WZB

Advanced natural language processing

- Advanced approaches from computational linguistics
  - *Part-of-Speech (POS) tagging:* nouns, verbs, adverbs, adjectives, ...
  - *Grammatical parsing:* e.g. subject-predicate-object structure
  - *Named entity recognition:* mark actors, institutions, places, etc. (often combined with large dictionaries / ontologies such as *JRC names* or *DBpedia*)
  - Note: Computationally expensive, not error free, and with varying quality across languages (but consistently improving)
- Very promising for political science analysis
  - Term disambiguation
  - More targeted identification of coding units
  - Event and interaction detection from political text
  - More accessible software tools looming (*spacy* and *spacyR*, e.g.)

WZB

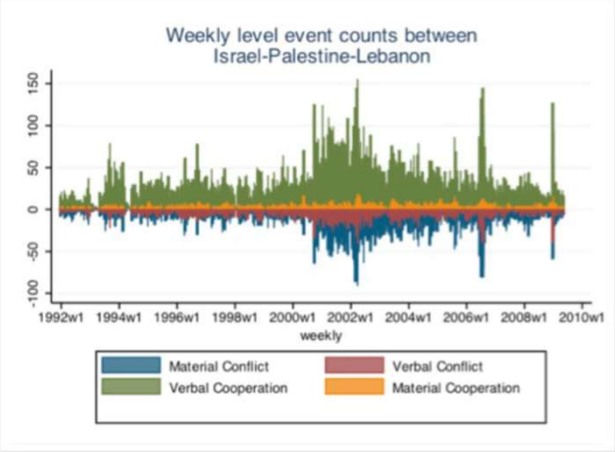
Example: Kansas Event Data System – KEDS (Schrodt / Gerner 1994)

- English language newswires (initially: first sentence of Reuters messages)
- Grammatical parsing
- Named entity recognition and verb categorization along very, very encompassing dictionaries
- Various 'debugging' routines for dealing with all sorts of vagaries of English language (passive voice, multiple objects, ...)
- Actor-Action-Actor data structure

Yesterday,	Lebanese forces	attacked	Israeli posts ....
	noun	*verb*	*noun*
	subject	*predicate*	*object*
	Dict: [Lebanon]	*Dict: [mat. conflict]*	*Dict.: [Israel]*

WZB

Example:
Parsed event data from newswires



Source: Presentation by P Schrodt (April 23 2013); visualization by J. Yonamine based on GDELT data

WZB

Seminar paper

- Goal: Research design proposal
  Study plan; own analyses not necessary (but also welcome!)
- Key elements the paper should contain
  - Research question of interest (IR, EU studies, political science)
    Very brief review of relevant literature and hypotheses
  - Justification for studying this question with text analysis
  - Possible sources, selection and preparation of relevant texts
  - Specification of the respective text analysis method(s) envisaged to retrieve systematic information from the texts
  - Discussion of insights to be gained and possible weaknesses
- Formalities
  - 4,000-5,000 words, English, pdf
  - Good scientific practice (originality, quoting and referencing)
  - Deadline: June 7 2019, midnight
  - Mail to: christian.rauh@wzb.eu

WZB

Promises and pitfalls of automated content analyses

+ A more complete and reliable analysis of social phenomena
  - Analysis of very large document sets achievable at low cost
  - Reduced / removed sampling bias

+/- Human resources remain significant
  - Dictionary development, coding of reference texts and especially validation requires intense human engagement

- Context dependency more pronounced
  - Quantitative representations of language cannot abstract from varying contexts (human coders can)

- Reliability is partially traded against validity
  - Power of automated analyses declines quickly with the complexity of theoretical concepts

WZB

Conclusions

- Automated text analyses are a powerful, yet not a definitive tool for content analysis in the Political and Social Sciences
- The computer allows us to digest larger amounts of information, uncovers patterns on much more aggregated levels, but interpretation, contextualisation, and validation remain key responsibility of the researcher!

Thank you for your attention!

Slides and tutorials available at www.christian-rauh.eu/teaching
[pw: CUP2019]